PR #43346 完整报告

vllm-project/vllm

[Metrics] Exclude KV transfer tokens from iteration_tokens_total

合并时间：2026-05-30 03:56

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43346

执行摘要

- 一句话：修复 iteration_tokens_total 包含 KV 传输 token 的问题
- 推荐动作：值得合入，改动精简且正确。建议在 release notes 中明确说明此变更，提醒已经在监控中使用该直方图的用户注意数值变化。

功能与动机

PR body 指出：当使用 P/D 分离部署时，external_kv_transfer token 会以每个新到达请求的完整 prompt 长度膨胀该指标，导致迭代 token 数远高于实际。使用 prompt_token_stats.computed 可排除 KV 传输和缓存命中的 token，使指标反映真实的本地计算量。

实现拆解

1. 定位问题所在：vllm/v1/metrics/loggers.py 中 histogram_iteration_tokens 的观测值使用 iteration_stats.num_prompt_tokens，该字段包含所有来源的 prompt token。
2. 修改观测值：将第 1164 行从 iteration_stats.num_prompt_tokens + iteration_stats.num_generation_tokens 改为 iteration_stats.prompt_token_stats.computed + iteration_stats.num_generation_tokens。
3. 一次先扩后缩的演进：第一版 commit 还尝试同步修改了 prompt_tokens 计数器，但经内部讨论后回退了该改动，仅保留直方图的修正，避免破坏现有监控面板的兼容性。
4. 无配套测试变更：本次改动仅 2 行源码，未添加或修改测试文件。

关键文件：

- vllm/v1/metrics/loggers.py（模块 指标；类别 source；类型 core-logic）：唯一的变更文件，修改了 histogram_iteration_tokens 的观测值，直接影响 Prometheus 指标输出。

关键符号：未识别

关键源码片段

vllm/v1/metrics/loggers.py

唯一的变更文件，修改了 histogram_iteration_tokens 的观测值，直接影响 Prometheus 指标输出。

```
# vllm/v1/metrics/loggers.py, 约 1162-1166 行
```

```
# 变更前：使用 total prompt tokens，包含 external_kv_transfer 导致指标虚高
# self.histogram_iteration_tokens[engine_idx].observe(
# iteration_stats.num_prompt_tokens + iteration_stats.num_generation_tokens
# )

# 变更后：仅使用 computed prompt tokens，与 LoggingStatLogger 一致
self.histogram_iteration_tokens[engine_idx].observe(
    iteration_stats.prompt_token_stats.computed
    + iteration_stats.num_generation_tokens
)

# 说明：computed 仅包含本地计算的 prompt token，
# 不包含 local_cache_hit 和 external_kv_transfer，
# 适用于 P/D 分离部署时准确反映每轮迭代的 token 处理量。
```

评论区精华

核心讨论：Reviewer njhill 指出，修改后 prompt_tokens_total（未带 label）会变得和带 label 的 prompt_tokens_total{source="local_compute"} 数值相同，可能造成混淆，且会破坏已有对此类问题做了修正的监控面板。作者 tlrmchlsmth 采纳建议，将改动范围缩小至仅修改直方图，保持计数器不变。 此外，自动代码审查机器人 gemini-code-assist 也指出了同样的不一致性问题，但作者已根据 njhill 的意见做了调整。

- 指标一致性与监控面板兼容性 (design): 作者将改动范围缩小至仅修改直方图，保持计数器不变，避免兼容性问题。

风险与影响

- 风险：风险较低：变更仅涉及 2 行核心逻辑，且仅修改了直方图的观测值，不影响任何控制流或数据路径。逆兼容性问题：对于已经在监控面板中手动使用 prompt_token_stats.computed 的用户，如果他们对 iteration_tokens_total 做了类似修正，升级后可能出现指标重报，但概率极小，且 PR body 已建议在 release notes 中声明。
- 影响：影响范围：仅影响 Prometheus 指标 iteration_tokens_total 直方图的输出值，对 P/D 分离部署用户影响较大，对单机部署用户无影响（因为 computed 与 num_prompt_tokens 在无 KV 传输时相等）。影响程度：指标准确性提升，预期不会引起功能回归或性能变化。
- 风险标记：缺少测试覆盖，指标兼容性影响

关联脉络

- 暂无明显关联 PR